2020

Model-Based Cluster Analysis of Indiana Social Security Beneficiary Data

Gwendolyn Spencer
Butler University

Model-Based Cluster Analysis of Indiana Social Security Beneficiary Data

Gwendolyn R. Spencer

April 29, 2020

An Undergraduate Thesis

Presented to The Honors Program

of Butler University

Supervised by Dr. Rasitha Jayasekare

In Partial Fulfillment

of the Requirements for Graduation Honors

Butler University Honors Program
Honors Thesis Certification

Applicant	Gwendolyn Spencer
Thesis Title	Model-Based Cluster Analysis of Indiana Social Security Beneficiary Data
Intended Date of Commencement	May 9, 2020

Read, Approved, and Signed by:

Thesis Adviser: ..... Date .....

Reader: ..... Date .....

Certified by: ..... Date .....

For Honors Program use:

Level of Honors conferred: University _____

Departmental _____

Acknowledgement

Special thank you to the Center for High Achievement and Scholarly Engagement (CHASE) at Butler University and to the Honors Program and Dr. Jason Lantzer for supporting this thesis and the presentation of this research at the 2020 Joint Mathematics Meeting (JMM) in Denver, Colorado. I would also like to thank the American Mathematical Society and the College of Liberal Arts and Sciences Dean at Butler University for supporting the presentation of this research at JMM.

Thank you to Dr. Chris Wilson for providing additional expertise and assistance in completing this paper. Additionally, thank you to the Department of Mathematics, Actuarial Science, and Statistics at Butler University for supporting this research, providing resources at the 2019 Mathematics Research Camp, and awarding myself the 2020 Judith Morrel Award, given to a student who achieves academic excellence in an honors project. Thank you to JMM for giving me the opportunity to present this research and for awarding me an honorable mention.

Lastly, many thanks to Dr. Rasitha Jayasekare, the advisor for this thesis, for her belief in this research, her constant support, and her dedication in the completion of this project. I am extremely grateful for all the help she has given me over the lifetime of this project, and know that none of this would have been possible without her.

ABSTRACT

Model-Based Cluster Analysis of Indiana Social Security Beneficiary Data

Gwendolyn Spencer

April 29, 2020

Annual reports of the U.S. Old-Age, Survivors, and Disability Insurance (OASDI) program, published by the Social Security Administration, detail the aggregate information about the program for each U.S. Postal ZIP code. This information includes the types of beneficiaries and monthly benefits received. These reports present the opportunity for contemporary analysis of the aggregate information about the OASDI program. To better capture the significance of the most-recent report for 2018, this project will use model-based cluster analysis, the unsupervised machine-learning process of grouping similar data points, to compare the 2017 and 2018 data. Due to the large amount of data, the project will look solely at the information for the state of Indiana. The form of model-based clustering used in this research assumes that the probability with which each data point belongs to a cluster is determined through a Gaussian mixture model. Maximum Likelihood Estimation will be used to estimate the parameters of the model and the Expectation-Maximization Algorithm will be used to complete this estimation. Bayesian Information Criterion will select the optimal number of clusters. This model should uncover underlying patterns in Social Security benefits paid in Indiana over recent years, as categorized by ZIP code.

Contents

List of Figures

List of Tables

Chapter 1

Background and Introduction

For many decades, citizens of the United States have appreciated the advantages of the Social Security Administration and the benefits paid out by the Old-Age, Survivors, and Disability Insurance program (OASDI). Despite the importance of the OASDI program in the United States, little research has been done to look at the patterns in benefits paid from the program. This research will examine the OASDI program reports, specifically looking at the most recent reports from 2017 and 2018 and solely at the data for the state of Indiana. To discover the underlying trends in the datasets, this research will use a popular data mining technique know as cluster analysis.

1.1 Methodology

Cluster analysis is a common form of unsupervised machine learning, which is particularly useful for identifying subgroups of a dataset in which the expected outcome for subgroups is unknown. There are no predetermined subgroups nor patterns for the Social Security data available, which is why cluster analysis is appropriate for this research. Because cluster analysis is an unsupervised machine learning technique and the dataset does not contain any predictor variables, an initial hypothesis is not necessary.

The goal of cluster analysis is to identify significant subgroups in the dataset where the data points within a cluster are similar to one another and are dissimilar from data points belonging to other clusters (Tan et. al. 2006). Cluster analysis is practical for discovering relationships between data points as well as patterns within the dataset. However, one of the main difficulties of cluster analysis is determining the clusters and what comprises a cluster. This research provides insight on the relationships of particular ZIP codes with regards to their number of Social Security beneficiaries and the total

value of their benefits.

While there are many possible algorithms available for cluster analysis, a Gaussian mixture model has been chosen to perform clustering for this research. Gaussian mixture models are preferred for clustering, over other models, due to their flexibility in shape, which allows the model to emulate multiple distribution shapes. The other main benefit of Gaussian mixture models is that each data point belongs to a cluster with a certain probability. In other clustering algorithms, points are allowed to be in just one cluster, while Gaussian mixture models allow for membership in multiple clusters. This overall flexibility of Gaussian mixture models should facilitate better results for the model and provide deeper insight into the underlying patterns in the data.

A mixture model is used when a set of data is comprised of natural subsets where each of them follow a different distribution (Klugman et. al 2012). In such cases an assumption of a singular distribution would be inappropriate. By assuming two or more distributions from two or more separate underlying regimes in the data, a mixture model is able to better capture the overall distribution of the data. For example, suppose heights of athletes, which consists of both males and females, are analyzed. It is possible that a situation may arise where the heights of females in the dataset follow a different distribution than the heights of males, in which case a mixture of distributions might be needed to model the heights of athletes. For mixture models, observations are assumed to be independent and identically distributed within each distribution.

In order to build the Gaussian Mixture Model, the parameters of the density function must be estimated. Maximum likelihood estimation has been chosen to estimate the parameters for this mixture model. Maximum likelihood estimation is a common method in statistics with the goal of estimating parameters that maximize the likelihood of producing the actual data. This method requires that a likelihood function be solved to find the parameter estimates. The likelihood function is the product of the density functions for each data point observed. A benefit of maximum likelihood estimation is that individual data points are used in estimating the parameters (Klugman et. al 2012). Since a mixture model is being used, each data point has a probability of belonging to multiple density functions. These probabilities are missing information that must be handled when estimating the parameters.

The most commonly used algorithm to handle missing information while estimating maximum likelihood parameters in mixture models is the Expectation and Maximization (EM) algorithm. The EM algorithm can find the maximum likelihood estimators if information is incomplete and can also be

used for a variety of models. For each data point, the EM algorithm initially assumes the probability that a point belongs to a cluster. It then computes the maximum likelihood parameter estimates using that assumption. After this, the algorithm recomputes the probability value and repeats this process until it converges on a value.

1.2 Literature Review

Gaussian mixture models have been applied across a wide range of fields. One common application of Gaussian mixture models is for use in speech recognition software. Povey et al. (2011) specifically explored using a Gaussian mixture model in speech recognition software in their paper, "The Subspace Gaussian Mixture Model—A Structured Model for Speech Recognition". A Gaussian mixture model was used for each Hidden Markov Model in the speech recognition software, giving better results than the conventional models typically used for similar speech recognition software programs.

Gaussian mixture models have also found applications in research for video surveillance. In the paper, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection," KaewTraKulPong and Bowden (2002) were able to use the Gaussian mixture model to help distinguish between moving shadows and moving objects in surveillance footage. The authors also saved time and space using the EM Algorithm for building their Gaussian mixture model, which further supports the decision to use the EM Algorithm in this research. The applications of the Gaussian mixture model in both speech recognition software and in video surveillance are important to consider, because it is useful to examine how the model can be applied to a wide range of datasets.

Beyond use in Gaussian mixture models, the EM algorithm has been utilized in the actuarial field for loss modeling and risk evaluation. Miljkovic and Gruen (2016) explored multimodality with distributions commonly used in actuarial loss modeling in their paper, "Modeling Loss Data using Mixtures of Distributions." The authors made an estimation of the models with a method based on the EM algorithm, which agains supports the use of the EM algorithm with this type of research.

The Social Security Administration and the OASDI program have also been of interest to many researchers. In *The Handbook of Public Economics*, Feldstein and Liebman (2002) discuss theoretical economic issues with Social Security and the risks associated with the Social Security plans they propose in their research. The authors discuss the Social Security Administration once again in their 2002 book, *The Distributional Aspects of Social Security*

and Social Security Reform. In their research, the authors explore the impact on Social Security as a result of the increasing life expectancy in the United States and the resulting impact on other retirement revenue sources. Also examining the economical side of Social Security, Diamond (1977) analyzed the provision of insurance to individuals that are unable to purchase insurance through the private insurance market sector and the nature of individuals to save their money in the paper, "A Framework for Social Security Analysis." This economic research of Social Security is necessary for considering the larger impact of the findings in the research.

In the paper, "Outcome Variation in the Social Security Disability Insurance Program: The Role of Primary Diagnoses" written for the Social Security Bulletin, Meseguer (2013) was able to conduct particularly intriguing research regarding the OASDI program. The goal of the author's research was to investigate the disability income portion of the OASDI program, specifically to see if medical diagnoses and state of origin afffected the disability decision that an individual received. This is especially interesting for the research being conducted as part of this research, because the author used cluster analysis to examine a portion of the OASDI program by state using data from 1997 - 2004.

Cluster analysis of economic data for specific geographical boundaries has been done before, as in the research conducted by Ahlquist and Breunig (2009) in their paper, "Country Clustering in Comparative Political Economy". The authors attempt to cluster countries based on their economies, with the goal of expanding the use of mixture models in the field of social sciences. In their research, the authors chose model-based clustering, specifically with a Gaussian mixture model, because the model selects the appropriate number of clusters according to the data, rather than relying on user input. The research in this paper similarly hopes to utilize the power of model-based clustering for finding patterns in social science data.

Gough (2001) also used cluster analysis to examine social science data in his paper, "Social Assistance Regimes: a Cluster Analysis". The author used two popular forms of cluster analysis, k-means clustering and hierarchical clustering, to find patterns in the social assistance programs of various countries. Though the author used a different form of cluster analysis than what will be used in this research, he was able to determine similar social assistance programs based on quantitative data.

1.3 Chapter Outline

The organization of the rest of this report is as follows. In Chapter 2, the dataset will be discussed and will be analyzed through outlier analysis. Chapter 3 discusses key concepts about the Gaussian Mixture Model, maximum likelihood estimation, and the expectation maximization algorithm. Chapter 4 explores the cluster analysis used, along with the cluster selection process and dimension reduction. In Chapter 5, the results of Chapter 4 will be discussed and the conclusions of this research will be revealed.

Chapter 2

Dataset Overview

For this research, both the 2017 ("OASDI Beneficiaries by State and ZIP Code, 2017") and 2018 ("OASDI Beneficiaries by State and ZIP Code, 2018") Old-Age, Survivors, and Disability Insurance (OASDI) program reports from the Social Security Administration will be used. The OASDI program is more commonly known as the U.S. Social Security program. These datasets consist of the types of beneficiaries and the amount of benefits paid through the OASDI program. The datasets are part of an annual publication of data from the Social Security Administration (SSA). Social insurance programs, like the OASDI program, exist in many countries to provide insurance to a larger group of people than those who are able to purchase insurance through the private insurance market.

The dataset is housed on data.gov, a site that holds all of the U.S. Government's public data. Data is usually submitted to the website by each government agency and the OASDI dataset was submitted to the website directly by the Social Security Administration.

The dataset shows the number of beneficiaries with benefits in current-payment status and total monthly benefits for December 2017 and December 2018. The main variables for the dataset are ZIP code, SSA field office, the number of retired workers, disabled workers, widow(er)s and parents, spouses and children that received beneficiary payments from OASDI, total monthly benefits for retired workers and widow(er)s, and the number of OASDI beneficiaries aged 65 or older. Disclosures of the amounts of benefits and the reasons for Social Security eligibility were avoided through the use of a controlled rounding procedure for the dataset entries. Each record in the data represents a singular ZIP code in the United States, except for ZIP codes in which there were less than 15 beneficiaries. To allow for more in-depth post-analysis in this research, this project will only use the ZIP codes located in the state of Indiana.

2.1 Data Summary

The variable, SSA Field Office, has 27 values, including: Anderson, Auburn, Bloomington, Columbus, Crawfordsville, Danville, IL, Elkhart, Evansville, Fort Wayne, Gary, Hammond, Indianapolis, Indianapolis NE, Indianapolis NW, Kokomo, Lafayette, Madison, Marion, Merrillville, Michigan City, Muncie, New Albany, Richmond, South Bend, Terre Haute, Valparaiso, and Vincennes. These values represent the field offices that pay out social security benefits to nearby ZIP codes. All variables besides Field Office and ZIP code are quantitative variables.

In the OASDI program for 2017 and 2018, there were 61,903,360 and 62,906,222 total beneficiaries respectively. In Indiana alone, there were a total of 1,335,288 and 1,350,417 beneficiaries in 2017 and 2018 respectively. The majority of these beneficiaries are retired workers, comprising over 67% of the program in 2017 and 2018 in Indiana. The total amount of monthly benefits paid in December of 2017 in Indiana was $1.784 billion, compared to $1.877 billion in December of 2018.

Shown in Table 2.1 are the number of ZIP codes that had zero recipients of each type of beneficiary for 2017 and 2018. For both retired workers and disabled workers, each ZIP code in the state contained at least 1 beneficiary. However, widow(er)s, parents, spouses, and children beneficiaries were only represented in around 90% of the ZIP codes.

Variable	ZIP Codes for 2017	ZIP Codes for 2018
Number of Retired Workers	0	0
Number of Disabled Workers	0	0
Number of Widow(er)s & Parents	81	75
Number of Spouses	135	139
Number of Children	79	83

Table 2.1: ZIP Codes with No Beneficiaries

The retired workers benefit portion of the OASDI program is typically the most well-known part of the program. This is shown in the data, through the mean of the number of retired workers variables, as summarized in Table 2.2. The mean of the number of retired workers was 190.3 in 2017 and 190.3 in 2018. In comparison, the number of disabled workers had a mean of 92.09 in 2017 and 92.99 in 2018. The mean of these variables represents the average number of the specific types of beneficiaries in a ZIP code in Indiana. As the data shows, on average, more retired workers receive benefits than disabled workers.

Variable	2017 Mean	2018 Mean	2017 Std Dev	2018 Std Dev
Num. of Retired Workers	190.3	190.3	104.8099	107.9917
Num. of Disabled Workers	92.09	92.99	47.94559	48.55899
Total Benefits (in 1000s)	358.6	359.5	198.2997	202.7328
WP Benefits (in 1000s)	141.6	137.98	100.9868	98.84516

Table 2.2: Summary Statistics for Select Variables

Due to the wide range of individuals eligible for benefits, it would be expected that the variable, total monthly benefits, would have a large standard deviation. Total monthly benefits (in thousands of dollars) has a standard deviation of 198.2997 in 2017 and 202.7328 in 2018. For comparison, the standard deviation of widow(er)s and parents monthly benefits (in thousands of dollars) was 100.9868 in 2017 and 98.84516 in 2018.

2.2 Outlier Analysis

Prior to outlier removal, the 2017 dataset had 875 records and the 2018 dataset had 872 records. For the variables, widow(er)s and parents, spouses, and children, the dataset for both years is right-skewed, as shown in figure 2.1 and figure 2.2. This is further shown by the measure of skewness for the variables. The variable widow(er)s and parents had a skewness of 2.052594 in 2017 and a skewness of 2.064878. Skewness values greater than positive 1 indicate variables that are highly skewed.

Spouses and children show similar levels of skewness to widow(er)s and parents, as summarized in Table 2.3. For the variable spouses, it had a skewness of 2.270863 in 2017 and 2.353922 in 2018. The skewness of the variable children was 2.144899 in 2017 and 2.102908 in 2018. Because of the presence of highly skewed variables, outliers must be carefully considered in this research.

Variable	Skewness for 2017	Skewness for 2018
Widow(er)s & Parents	2.052594	2.064878
Spouses	2.270863	2.353922
Children	2.144899	2.102908

Table 2.3: Skewness for Select Variables for 2017 and 2018

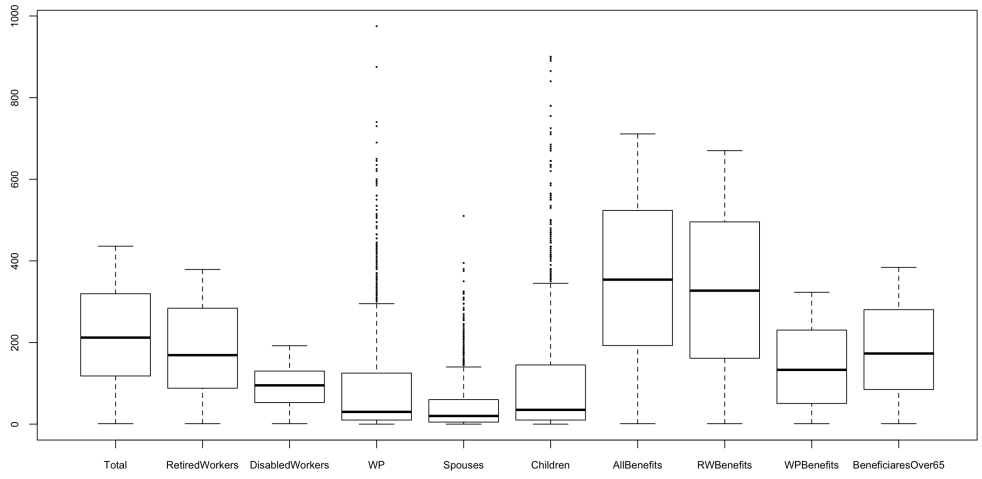

Figure 2.1: 2017 Boxplots

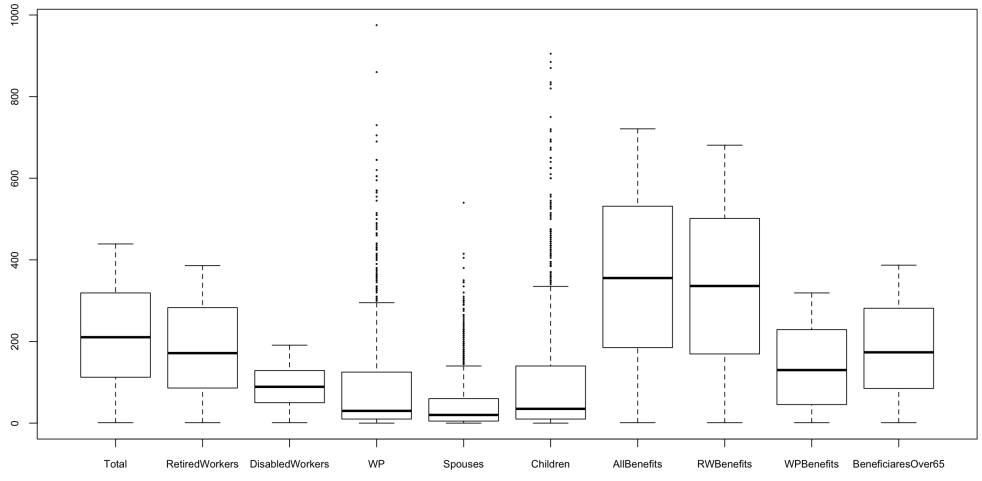

Figure 2.2: 2018 Boxplots

After outlier removal, the 2017 dataset contains 771 records and the 2018 dataset contains 770 records. As shown in Figure 2.3 and Figure 2.4, the boxplots are significantly less skewed after removing outliers.

Since outliers exist in the dataset, the clustering will be performed both with outliers included and outliers removed to determine which dataset is

optimal for clustering. It is important to consider both options, since the datasets represent the entire state of Indiana and it would be less wholistic to simply remove outliers from the dataset.

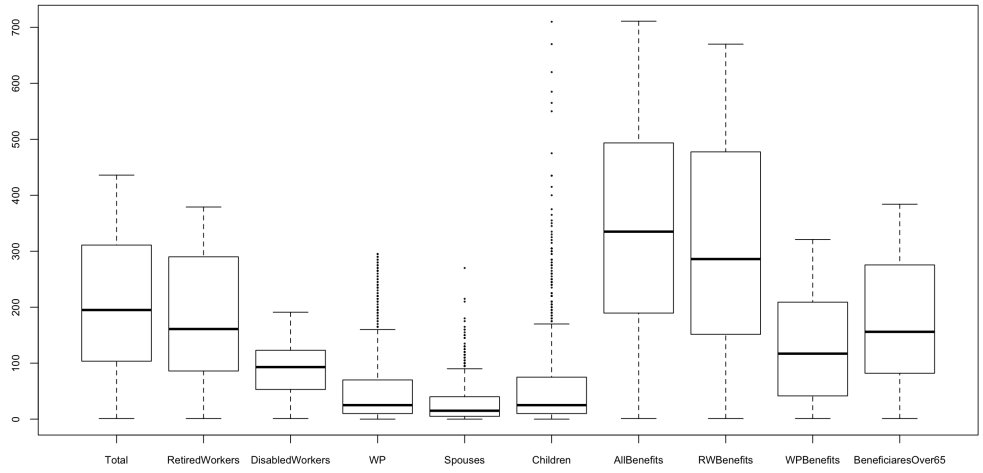

Figure 2.3: 2017 Boxplots after Outlier Removal

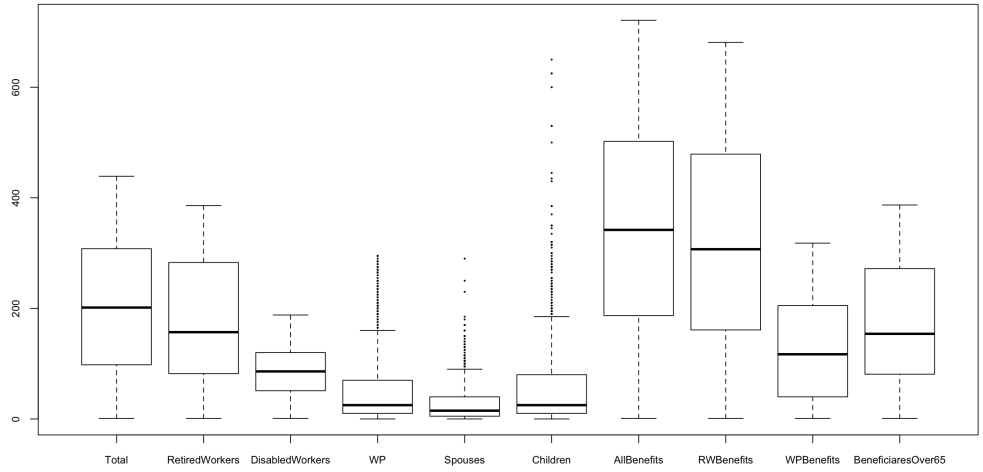

Figure 2.4: 2018 Boxplots after Outlier Removal

Chapter 3

Gaussian Mixture Model

The form of model-based clustering used in this research assumes the probability that any particular data point belongs to a cluster is determined through a Gaussian Mixture Model. A mixture model allows each cluster to be represented by a unique probability distribution, with each proability distribution specifically being a Gaussian distribution. In order to estimate the parameters of the model, maximum likelihood estimation is used. However, the expectation maximization algorithm must also be used to estimate the probability that any particular data point belongs to a cluster.

3.1 Maximum Likelihood Estimation

Though a Gaussian Mixture Model will be used for clutser analysis, the parameters of the model are unknown. Maximum likelihood estimation is a very common method for estimating parameters in a model. Maximum likelihood estimation is often prefered, because it uses the individual data points to estimate the parameters. The parameter estimates produced through this process maximize the likelihood that the model produces the actual data.

The likelihood function is the product of the density functions for each data point observed. When finding the maximum likelihood estimates, it is often easier to maximize the log-likelihood function, which is simply the natural logarithm of the likelihood fuction. To maximize the log-likelihood function in this research, derivatives are used to determine maximum values.

The likelihood function for the model is as follows:

Given observations $y = (y_1, \ldots, y_n)$, let $f_k(y_i|\theta_k)$ be the density of an observation y_i from the k^{th} component, θ_k be the corresponding parameter in the k^{th} component, and G be the number of components in the mixture.

Then, the likelihood of the mixtures is:

$$L_M(\theta_1, \ldots, \theta_G; \Delta_1, \ldots, \Delta_G | y) = \Pi_{i=1}^{n} \Sigma_{k=1}^{G} \Delta_k f_k(y_i|\theta_k)$$

where Δ_k is the probability that an observation belongs to the k^{th} component ($\Delta_k \geq 0;\ \Sigma_{k=1}^{G} \Delta_k = 1$)

3.1.1 Derivation of Maximum Likelihood Estimates in a 2 Component Gaussian Mixture Model

To generalize the derivation of the maximum likelihood estimates of a Gaussian mixture model, the proof below assumes only a 2 component model. First, given observations $y_1, \ldots, y_n$, let $\theta = \{p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$, where $\{p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$ are the parameters of the model. The likelihood function of the 2 component model can be written as:

$$L(y_1, \ldots, y_n, \theta) = \prod_{i=1}^{n} \Big((1-p) f_{y_1}(y_i)\Big)^{1-\Delta_i} \Big((p) f_{y_2}(y_i)\Big)^{\Delta_i}$$

where Δ_i represents the probability that a data point belongs to a component and can take the value of 0 or 1. Under the assumption of a Gaussian model, $f_{y_1}(y_i)$ and $f_{y_2}(y_i)$ are given by:

$$f_{y_1}(y_i) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}}$$

$$f_{y_2}(y_i) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}\frac{(y_i-\mu_2)^2}{\sigma_2^2}}$$

Taking the natural logarithm of both sides of the likelihood function and

simplifying results in the log-likelihood function:

$$
\begin{aligned}
\ln L(\theta) &= \ln\Bigg(\prod_{i=1}^{n}\Big((1-p)f_{y_1}(y_i)\Big)^{1-\Delta_i}\Big((p)f_{y_2}(y_i)\Big)^{\Delta_i}\Bigg) \\
&= \sum_{i=1}^{n}\ln\Bigg(\Big((1-p)f_{y_1}(y_i)\Big)^{1-\Delta_i}\Big((p)f_{y_2}(y_i)\Big)^{\Delta_i}\Bigg) \\
&= \sum_{i=1}^{n}\Bigg((1-\Delta_i)\ln\Big((1-p)f_{y_1}(y_i)\Big) + \Delta_i\ln\Big((p)f_{y_2}(y_i)\Big)\Bigg) \\
&= \sum_{i=1}^{n}\Bigg((1-\Delta_i)\ln(1-p) + (1-\Delta_i)\ln\Big(f_{y_1}(y_i)\Big) + \Delta_i\ln(p) + \Delta_i\ln\Big(f_{y_2}(y_i)\Big)\Bigg) \\
&= \sum_{i=1}^{n}\Bigg((1-\Delta_i)\ln(1-p) + (1-\Delta_i)\ln\Big(\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}}\Big) + \Delta_i\ln(p) + \\
&\quad \Delta_i\ln\Big(\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{1}{2}\frac{(y_i-\mu_2)^2}{\sigma_2^2}}\Big)\Bigg) \\
&= \sum_{i=1}^{n}\Bigg((1-\Delta_i)\ln(1-p) + (1-\Delta_i)\Big(-\frac{1}{2}\ln(2\pi\sigma_1^2) - \frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}\Big) + \Delta_i\ln(p) + \\
&\quad \Delta_i\Big(-\frac{1}{2}\ln(2\pi\sigma_2^2) - \frac{1}{2}\frac{(y_i-\mu_2)^2}{\sigma_2^2}\Big)\Bigg)
\end{aligned}
\tag{3.1}
$$

Then, each parameter can be estimated through maximum likelihood

estimation. To begin, the parameter, p, will be estimated:

$$\begin{aligned}
\frac{d\ln L(\theta)}{dp} &= \sum_{i=1}^{n}\left((1-\Delta_i)\frac{1}{1-p}(-1)+\frac{\Delta_i}{p}\right)\\
0 &= \sum_{i=1}^{n}\left((1-\Delta_i)\frac{1}{1-p}(-1)+\frac{\Delta_i}{p}\right)\\
\sum_{i=1}^{n}\left((1-\Delta_i)\frac{1}{1-p}\right) &= \sum_{i=1}^{n}\left(\frac{\Delta_i}{p}\right)\\
\sum_{i=1}^{n}\Big((1-\Delta_i)p\Big) &= \sum_{i=1}^{n}\Big(\Delta_i(1-p)\Big)\\
p\sum_{i=1}^{n}1-p\sum_{i=1}^{n}\Delta_i &= \sum_{i=1}^{n}\Delta_i - p\sum_{i=1}^{n}\Delta_i\\
pn &= \sum_{i=1}^{n}\Delta_i\\
\hat{p} &= \frac{\sum_{i=1}^{n}\Delta_i}{n}
\end{aligned} \tag{3.2}$$

Next, the parameters, μ_1 and μ_2, will be estimated:

$$\begin{aligned}
\frac{d\ln L(\theta)}{d\mu_1} &= \sum_{i=1}^{n}\left((1-\Delta_i)\Big(-\frac{1}{2\sigma_1^2}2(y_i-\mu_1)(-1)\Big)\right)\\
0 &= \sum_{i=1}^{n}\left((1-\Delta_i)\Big(-\frac{1}{2\sigma_1^2}2(y_i-\mu_1)(-1)\Big)\right)\\
0 &= \sum_{i=1}^{n}\left((1-\Delta_i)\Big(\frac{y_i-\mu_1}{\sigma_1^2}\Big)\right)\\
0 &= \sum_{i=1}^{n}\Big(\frac{(1-\Delta_i)y_i}{\sigma_1^2}\Big)-\sum_{i=1}^{n}\Big(\frac{(1-\Delta_i)\mu_1}{\sigma_1^2}\Big)\\
\sum_{i=1}^{n}\Big((1-\Delta_i)y_i\Big) &= \sum_{i=1}^{n}\Big((1-\Delta_i)\mu_1\Big)\\
\hat{\mu}_1 &= \frac{\sum_{i=1}^{n}\Big((1-\Delta_i)y_i\Big)}{\sum_{i=1}^{n}\Big((1-\Delta_i)\Big)}
\end{aligned} \tag{3.3}$$

$$
\begin{aligned}
\frac{d\ln L(\theta)}{d\mu_2} &= \sum_{i=1}^{n}\Big(\Delta_i\Big(-\frac{1}{2\sigma_2^2}2(y_i-\mu_2)(-1)\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big(\Delta_i\Big(-\frac{1}{2\sigma_2^2}2(y_i-\mu_2)(-1)\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big(\Delta_i\Big(\frac{y_i-\mu_2}{\sigma_2^2}\Big)\Big)\\
\sum_{i=1}^{n}\Big(\frac{\Delta_i y_i}{\sigma_2^2}\Big) &= \sum_{i=1}^{n}\Big(\frac{\Delta_i \mu_2}{\sigma_2^2}\Big)\\
\sum_{i=1}^{n}(\Delta_i y_i) &= \sum_{i=1}^{n}(\Delta_i \mu_2)\\
\hat{\mu}_2 &= \frac{\sum_{i=1}^{n}(\Delta_i y_i)}{\sum_{i=1}^{n}(\Delta_i)}
\end{aligned}
\tag{3.4}
$$

Lastly, the parameters, σ_1^2 and σ_2^2, will be estimated:

$$
\begin{aligned}
\frac{d\ln L(\theta)}{d\sigma_1^2} &= \sum_{i=1}^{n}\Big((1-\Delta_i)\Big(-\frac{1}{2}\Big(\frac{1}{2\pi\sigma_1^2}\Big)2\pi-\frac{1}{2}\frac{(y_i-\mu_1)^2}{(\sigma_1^2)^2}(-1)\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big((1-\Delta_i)\Big(-\frac{1}{2}\Big(\frac{1}{2\pi\sigma_1^2}\Big)2\pi-\frac{1}{2}\frac{(y_i-\mu_1)^2}{(\sigma_1^2)^2}(-1)\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big((1-\Delta_i)\Big(-\frac{1}{2\sigma_1^2}+\frac{(y_i-\mu_1)^2}{2\sigma_1^4}\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big((1-\Delta_i)\Big(-1+\frac{(y_i-\mu_1)^2}{\sigma_1^2}\Big)\Big)\\
\sum_{i=1}^{n}(1-\Delta_i) &= \sum_{i=1}^{n}\frac{(1-\Delta_i)(y_i-\mu_1)^2}{\sigma_1^2}\\
\hat{\sigma}_1^2 &= \frac{\sum_{i=1}^{n}(1-\Delta_i)(y_i-\mu_1)^2}{\sum_{i=1}^{n}(1-\Delta_i)}
\end{aligned}
\tag{3.5}
$$

$$
\begin{aligned}
\frac{d\ln L(\theta)}{d\sigma_2^2} &= \sum_{i=1}^{n}\Big(\Delta_i\Big(-\frac{1}{2}\Big(\frac{1}{2\pi\sigma_2^2}\Big)2\pi - \frac{1}{2}\frac{(y_i-\mu_2)^2}{(\sigma_2^2)^2}(-1)\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big(\Delta_i\Big(-\frac{1}{2}\Big(\frac{1}{2\pi\sigma_2^2}\Big)2\pi - \frac{1}{2}\frac{(y_i-\mu_2)^2}{(\sigma_2^2)^2}(-1)\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big(\Delta_i\Big(-\frac{1}{2\sigma_2^2} + \frac{(y_i-\mu_2)^2}{2\sigma_2^4}\Big)\Big)\\
0 &= \sum_{i=1}^{n}\Big(\Delta_i\Big(-1 + \frac{(y_i-\mu_2)^2}{\sigma_2^2}\Big)\Big)\\
\sum_{i=1}^{n}\Delta_i &= \sum_{i=1}^{n}\frac{\Delta_i(y_i-\mu_2)^2}{\sigma_2^2}\\
\hat{\sigma}_2^2 &= \frac{\sum_{i=1}^{n}\Delta_i(y_i-\mu_2)^2}{\sum_{i=1}^{n}\Delta_i}
\end{aligned}
\tag{3.6}
$$

In Table 3.1, the resulting maximum likelihood estimators are shown. Though the models used in this research will most likely use more than 2 components, a derivation similar to that above can be performed for models with greater than 2 components. The process to determine the optimal number of components for the models used in this research will be discussed in chapter 4.

$$
\begin{aligned}
\hat{p} &= \frac{\sum_{i=1}^{n}\Delta_i}{n}\\
\hat{\mu}_1 &= \frac{\sum_{i=1}^{n}\Big((1-\Delta_i)y_i\Big)}{\sum_{i=1}^{n}\Big((1-\Delta_i)\Big)}\\
\hat{\mu}_2 &= \frac{\sum_{i=1}^{n}\Big(\Delta_i y_i\Big)}{\sum_{i=1}^{n}\Big(\Delta_i\Big)}\\
\hat{\sigma}_1^2 &= \frac{\sum_{i=1}^{n}(1-\Delta_i)(y_i-\mu_1)^2}{\sum_{i=1}^{n}\Big(1-\Delta_i\Big)}\\
\hat{\sigma}_2^2 &= \frac{\sum_{i=1}^{n}\Delta_i(y_i-\mu_2)^2}{\sum_{i=1}^{n}\Big(\Delta_i\Big)}
\end{aligned}
$$

Table 3.1: Maximum Likelihood Estimators

3.2 Expectation Maximization Algorithm

When finding the maximum likelihood parameter estimates for the Gaussian Mixture Model, it is assumed that the probability that each data point belongs to a particular cluster is known. However, when applying a Gaussian

Mixture Model to a dataset for clutser analysis, these probabilities are unknown and must be estimated in order to complete the maximum likelihood parameter estimates. One of the most popular techniques for estimating these probabilities is known as the Expectation Maximization (EM) algorithm.

Though the EM algorithm is a popular technique for cluster analysis, it does have some limitations. Notably, the EM algorithm can be slow and can be impractical for models with a large number of clusters. However, because the size of the datasets used with the models in this research are small and the optimal number of clusters should also be small, the EM algorithm will be appropriate for this research.

The EM algorithm is a general approach to maximum likelihood estimation when the probabilities that a data point belongs to a particular cluster are unknown. EM algorithm initially assumes that a particular data point belongs to one of the clusters and calculates the maximum likelihood estimates with that assumption through the M-step. Then the algorithm solves for the probability that the data point belongs to that cluster with the calcualted parameter estimates for the E-step. Finally, the algorithm repeats the M-step and E-step until it converges on a specific value and satisfies convergance criteria. The EM algorithm repeats these steps for all data points and all clusters.

In EM algorithm, complete data is represented by:

$$y_i = (x_i, z_i) \text{ where } z_i = (z_{i1}, \ldots, z_{iG}) \text{ with } z_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to group } k \\ 0 & \text{otherwise} \end{cases}$$

The EM algorithm steps are as follows (Fraley and Raftery 1998):

initialize $\hat{z}_{ik}$
repeat
M-step: compute ML parameter estimates given $\hat{z}_{ik}$

$$n_k \leftarrow \Sigma_{i=1}^{n} \hat{z}_{ik}$$
$$\hat{\tau}_k \leftarrow \frac{n_k}{n}$$
$$\hat{\mu}_k \leftarrow \frac{\Sigma_{i=1}^{n} \hat{z}_{ik} x_i}{n_k}$$
$$\hat{\Sigma}_k : \text{depends on the model}$$

E-step: compute $\hat{z}_{ik}$ using the M-step

$$\hat{z}_{ik} \leftarrow \frac{\hat{\tau}_k f_k(x_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\Sigma_{j=1}^{G} \hat{\tau}_j f_j(x_i|\hat{\mu}_j, \hat{\Sigma}_j)}$$

$$f_k(x_i|\mu_k, \Sigma_k) = \frac{exp\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\}}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}$$

until convergence criteria are satisfied

Through the use of both the EM algorithm and maximum likelihood estimation, the parameters can be estimated for the Gaussian mixture model. This model will then be used to perform the clustering for both the 2017 and 2018 Indiana Social Security beneficiary datasets.

Chapter 4

Model-Based Cluster Analysis

After estimating the appropriate parameters for the Gaussian Mixture Model, as discussed in Chapter 3, the model-based cluster analysis of the datasets can be performed. However, in order to perform the clustering, the optimal number of clusters and dimension reduction must also be considered. This chapter will explore both cluster selection and dimension reduction.

4.1 Cluster Selection

A variety of measures exist to help determine the optimal number of clusters to use for a given dataset. In this research, the measure selected for this task is known as Bayesian Information Criterion (BIC). BIC is favorable when the EM algorithm is used to find the maximum likelihood estimates, making it appropriate for this research. To calculate the BIC for a model, the maximized mixture log-likehood function and the number of independent parameters in the model must be known.

For model-based clustering, the BIC formula used is:

$$\text{BIC} = 2l_{\mathcal{M}}(x, \hat{\theta}) - m_{\mathcal{M}} log(n)$$

where $l_{\mathcal{M}}(x, \hat{\theta})$: the maximized mixture log likelihood of the model $\mathcal{M}$,

$m_{\mathcal{M}}$: the number of independent parameters in the model

Model	BIC	Optimal Number of Clusters
2017	-2,116.457	9
2017 - No outliers	-3,157.63	8
2018	-2,291.491	8
2018 - No outliers	-3,133.478	9

Table 4.1: BIC Values for Models

4.1.1 BIC Graphs

Once the values for the BIC for all variations of models for a particular dataset have been calculated, the model with the highest BIC is selected. The model with the highest BIC has the strongest evidence for its use. As mentioned in Chapter 2, outliers must still be considered when developing the model, since there was not signficant evidence for proceeding with or without outliers in the model.

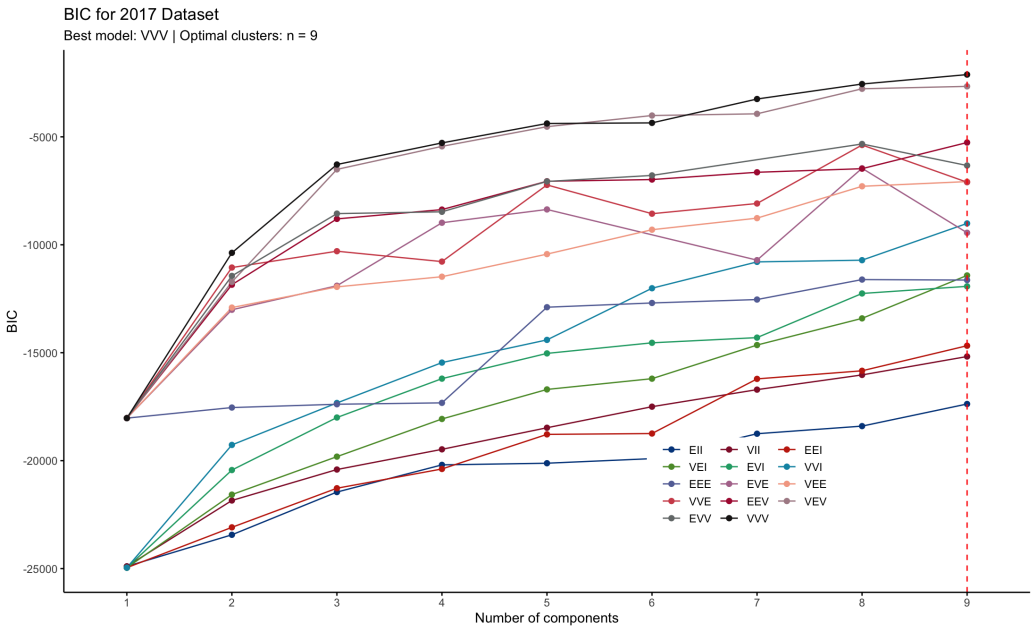

Figure 4.1: 2017 BIC Graph

Shown in Figures 4.1, 4.3, 4.2, and 4.4 are the BIC graphs for the 2017 and 2018 datasets, both before and after outlier removal. The summary of the graphs can also be found in Table 4.1. As shown in Figure 4.1, for the

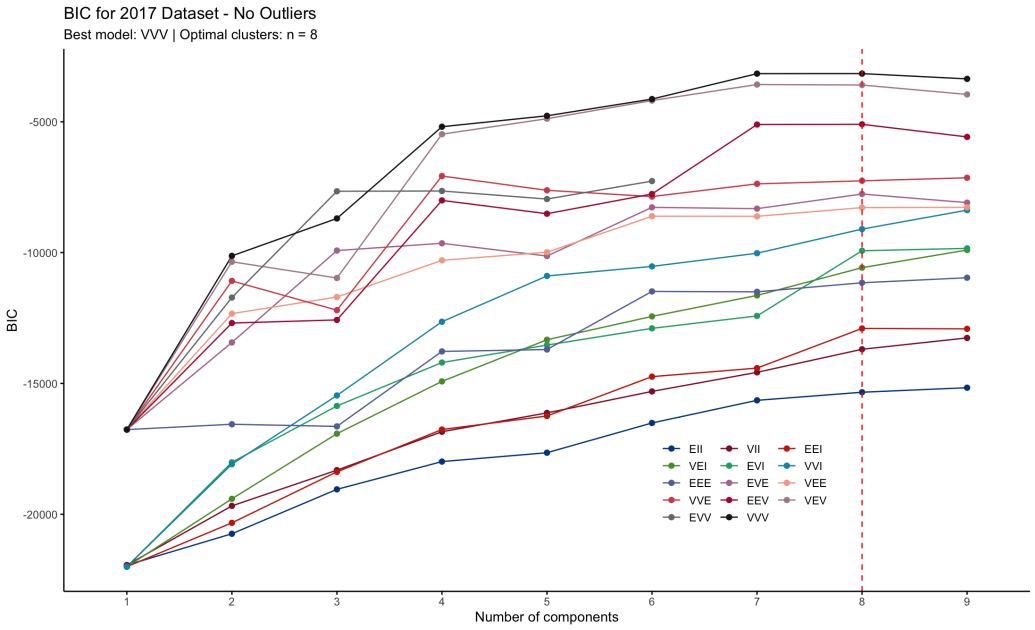

Figure 4.2: 2017 BIC Graph - No Outliers

2017 dataset prior to outlier removal, 9 clusters are determined to be the optimal number and the BIC for the optimal model is -2116.457. For the 2017 dataset after outlier removal, 8 clusters are determined to be the optimal number and the BIC for the optimal model is -3157.63, as shown in Figure 4.2. However, it is important to note that the BIC values of both years are not comparable since the BIC is specific to the set of data used in the model. Therefore, BIC will not be a determining factor in whether to consider the whole dataset or the dataset with outliers removed. Regardless, BIC does indicate the optimal number of clusters and is useful for that purpose.

From Figure 4.3, the optimal number of clusters for the 2018 data set (prior to outlier removal) is 8 clusters, and the BIC for the optimal model is -2291.491. For the 2018 dataset after outlier removal, the optimal number of clusters for the 2018 data set after outlier removal is 9 clusters, and the BIC for the optimal model is -3133.478 , as shown in Figure 4.4.

After running all four models, the optimal number of clusters varies between 8 and 9 clusters for the datasets. It is expected for the dataset to change over time since different individuals will qualify or elect to qualify for OASDI benefits each years. Also, the number of ZIP codes included in the OASDI program report changes over time, due to the omission of ZIP codes with under 15 beneficiaries in the program reports. Because the dataset is not consistent year to year, it is logical that the clustering algorithm would

find differing optimal numbers of clusters.

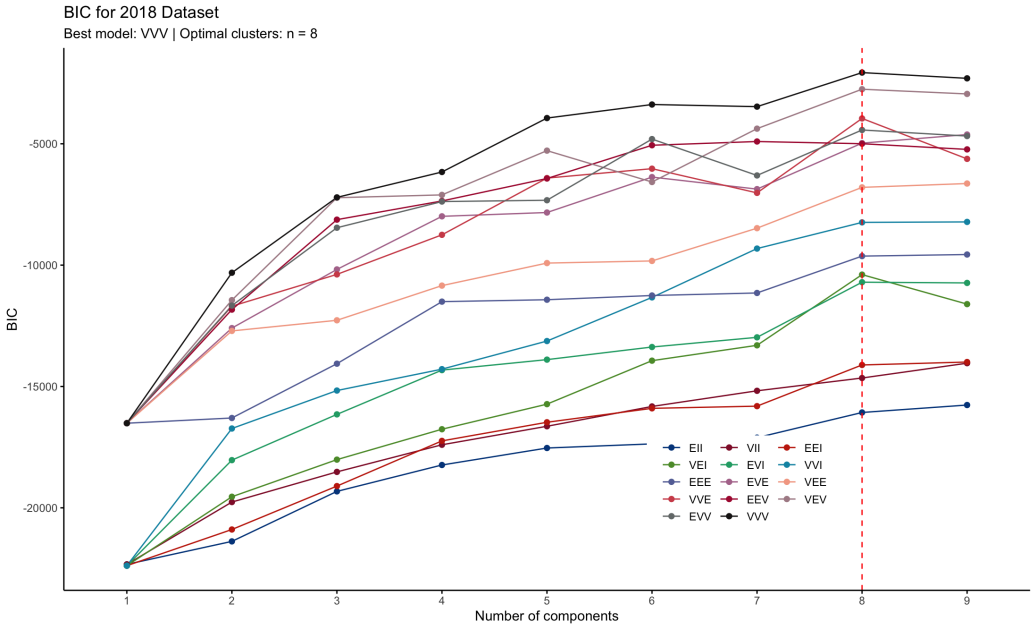

Figure 4.3: 2018 BIC Graph

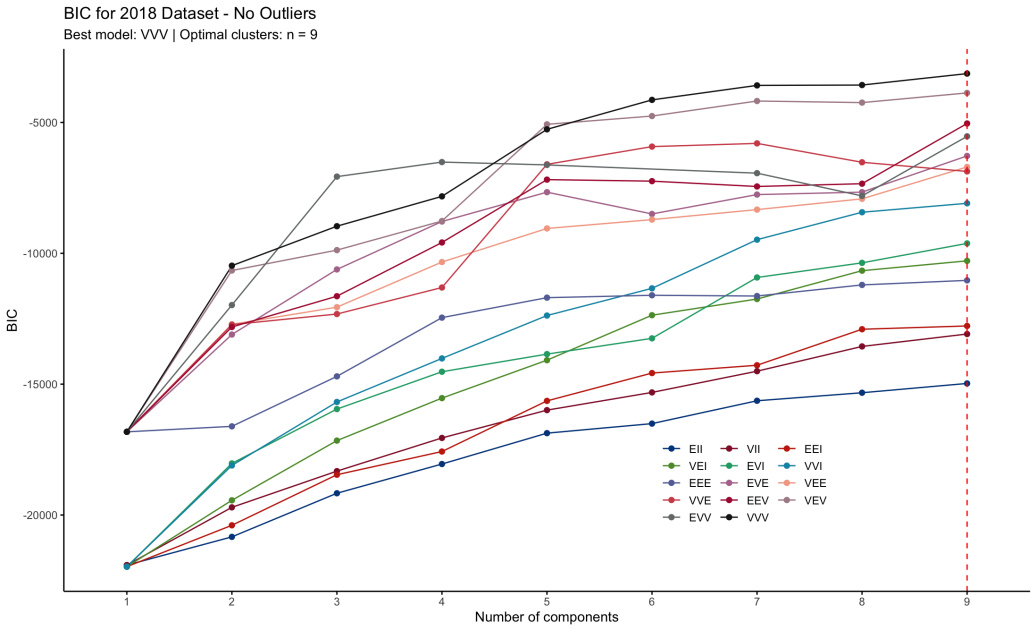

Figure 4.4: 2018 BIC Graph - No Outliers

Since both datasets have roughly 870 data points before outlier removal and 770 data points after outlier removal, either 8 or 9 clusters is an appropriate number. If too many clusters are selected, the clusters will contain very few data points, while if too few clusters are selected, the clusters will contain too many data points. In all the clusters for all four models, the smallest amount of data points contained in a cluster is 24, which is sufficiently large to provide meaningful information for further analysis.

4.2 Dimension Reduction

After selecting the appropriate number of clusters using BIC, it is logical to want to visualize the clusters found in the model. However, if the clusters are plotted at this stage, it would be complicated to interpret the graphs, because the clusters are difficult to visualize. To visualize the clusters, dimension reduction can be used.

There are many techniques commonly used in data mining for dimension reduction, including Principal Component Analysis. The goal of these techniques is to select the appriopriate number of variables that can explain the variability of the data in order to avoid the curse of dimensionality. However, a method proposed by Scrucca (2015) uses dimension reduction with model-based clustering for the purpose of visualizing the clustered data.

The goal of dimension reduction in this context is to find a subspace that still captures most of the clustering information in the data. Though dimension reduction is typically conducted as a data pre-processing step, it is used in this purpose for visualizing the clusters and is therefore appropriate to use after the clustering has been performed.

Shown in Figure 4.5, Figure 4.7, Figure 4.6, and Figure 4.8 are the density plots for the 2017 and 2018 datasets, both before and after outlier removal. One benefit of the density plots is the ability to capture to need for a mixture model to represent the datasets. As shown in each density plot, each cluster has a unique density curve and only overlaps in some cases.

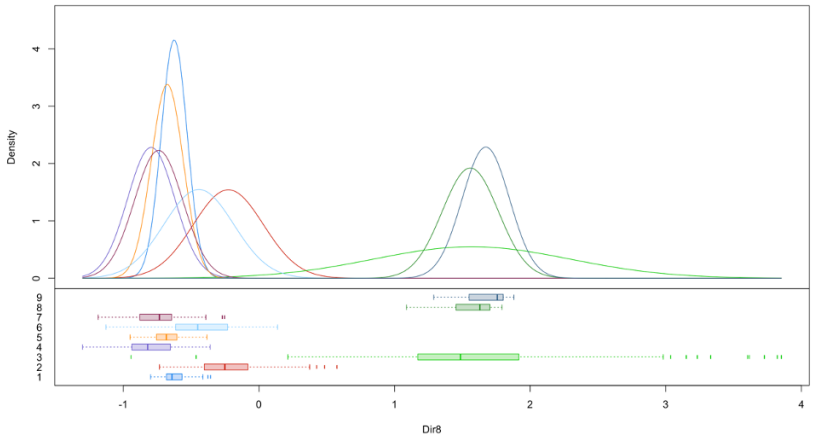

Figure 4.5: 2017 Density Plot

For the 2017 dataset, it is very noticeable how the clusters improve after outlier removal. In Figure 4.5, the density curves are overlapping and not clearly seperated, even at a high dimension. However, in figure 4.6, the density curves are more clearly separated and less overlapping. In the 2018 dataset, the change before and after outlier removal is less distinct.

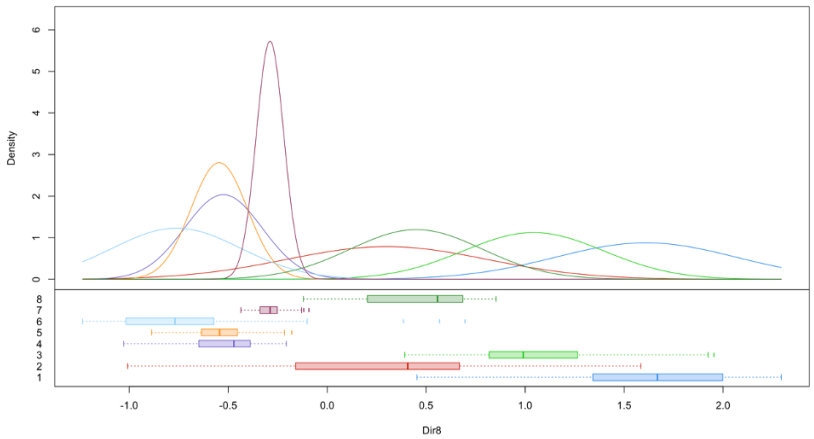

Figure 4.6: 2017 Density Plot - No Outliers

After considering all the models and whether or not to remove outliers, it is important to remember that the goal of this research is to find patterns in the OASDI benefits paid in the state of Indiana. Since the state of Indiana is comprised of the ZIP codes in the original datasets, it would be more wholistic to proceed with outliers in the datasets. As shown throughout this

chapter, there have only been minor improvements to the models for both years when outliers are removed. Therefore, it is acceptable to proceed with outliers in the datasets.

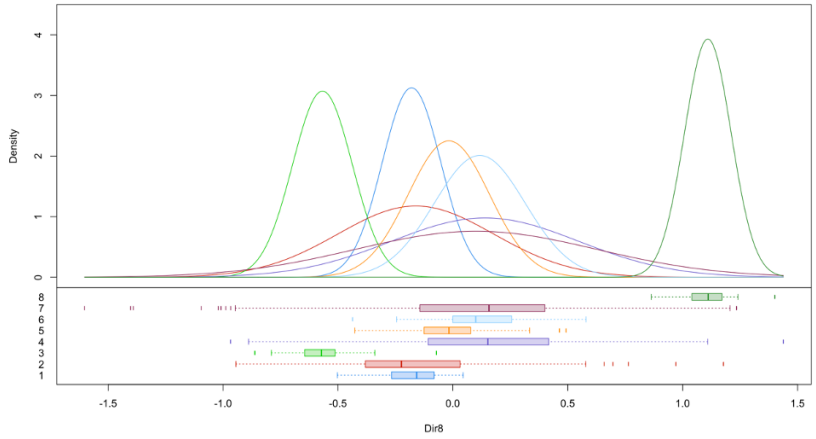

Figure 4.7: 2018 Density Plot

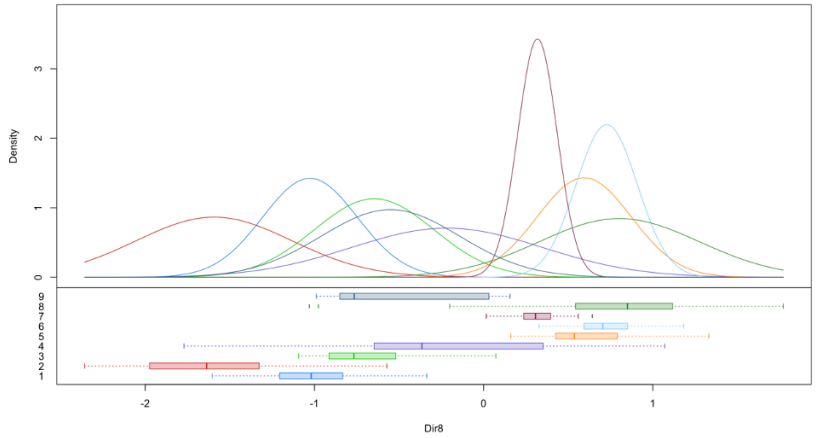

Figure 4.8: 2018 Density Plot - No Outliers

Chapter 5

Post-Analysis of Clustering

After successfully completing the cluster analysis, the results from the clustering and the implications of the models can be determined. As discussed in Chapter 4, the two models that were selected to be further analyzed were selected from the 2017 and 2018 OASDI reports, inclusive of outliers. This chapter will examine the patterns that exist in both datasets and between both models.

5.1 Results

In evaluating the clusters for the 2017 and 2018, it is necessary to compare the results of the clustering both to the data and to the external information. To first gain a better understanding of the clustering for both years, shown in Figure 5.1 and Figure 5.2 are maps of the ZIP codes for Indiana, color coded by the cluster each ZIP code belongs to for 2017 and 2018. This means that each ZIP code that is colored white in the 2017 map belongs to the same cluster and each ZIP that is colored dark red in the 2017 map belongs to the same cluster, etc.

From Figure 5.1 and Figure 5.2, it is possible to see how the clusters are distributed geographically. In both years, the ZIP codes in Indianapolis, Fort Wayne, Evansville, and South Bend belong mostly to one of two clusters. These are the four largest cities in Indiana, so it would be expected that their populations of individuals who would be eligible for OASDI benefits would be more similar than the ZIP codes of more rural areas.

Similarly, the four largest universities in Indiana are Purdue University, Indiana University, Indiana University Purdue University Indianapolis, and Ball State University, which are located in West Lafayette, Bloomington, Indianapolis and Muncie respectively. As seen in the map, the ZIP codes in

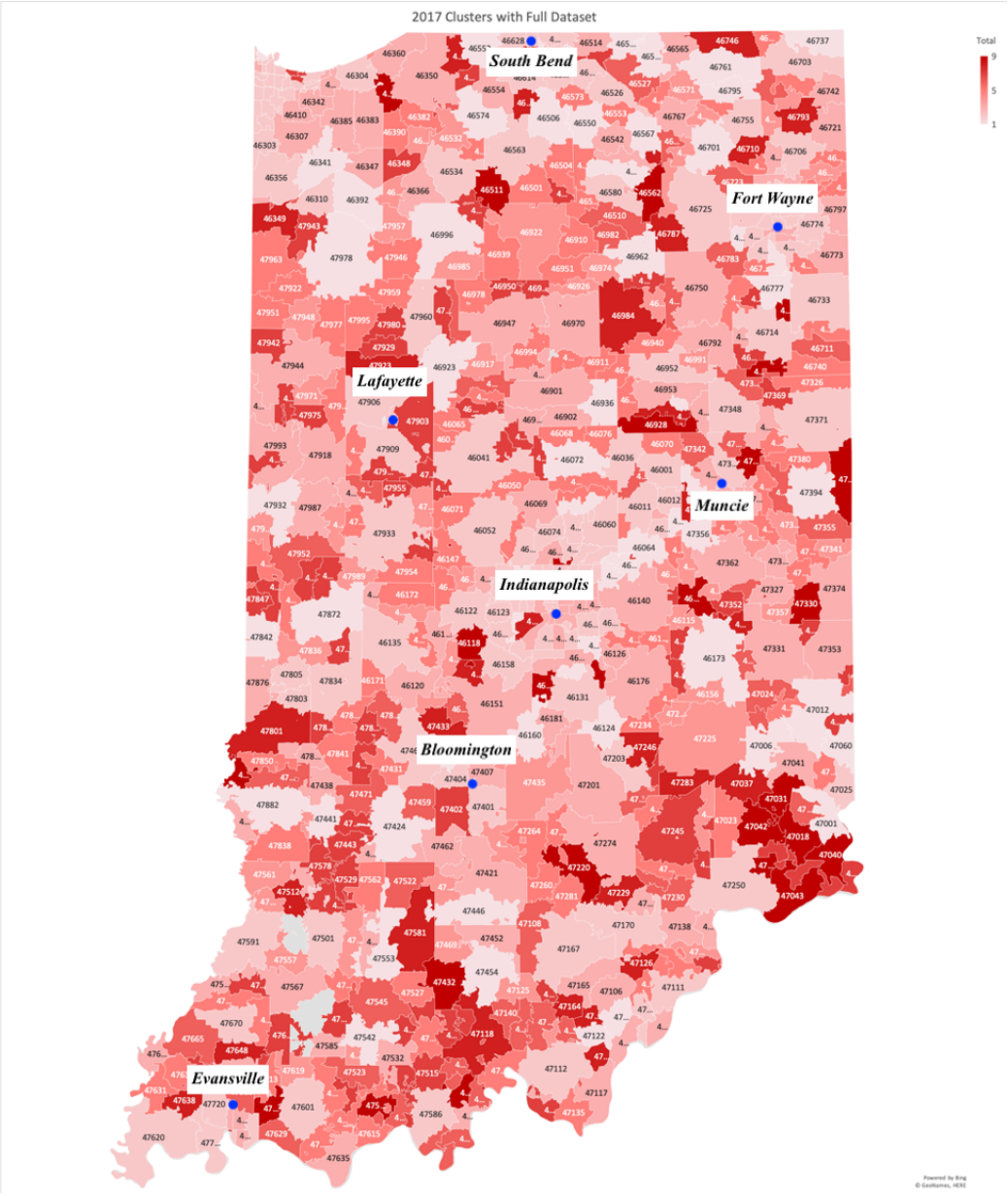

Figure 5.1: 2017 Cluster Map

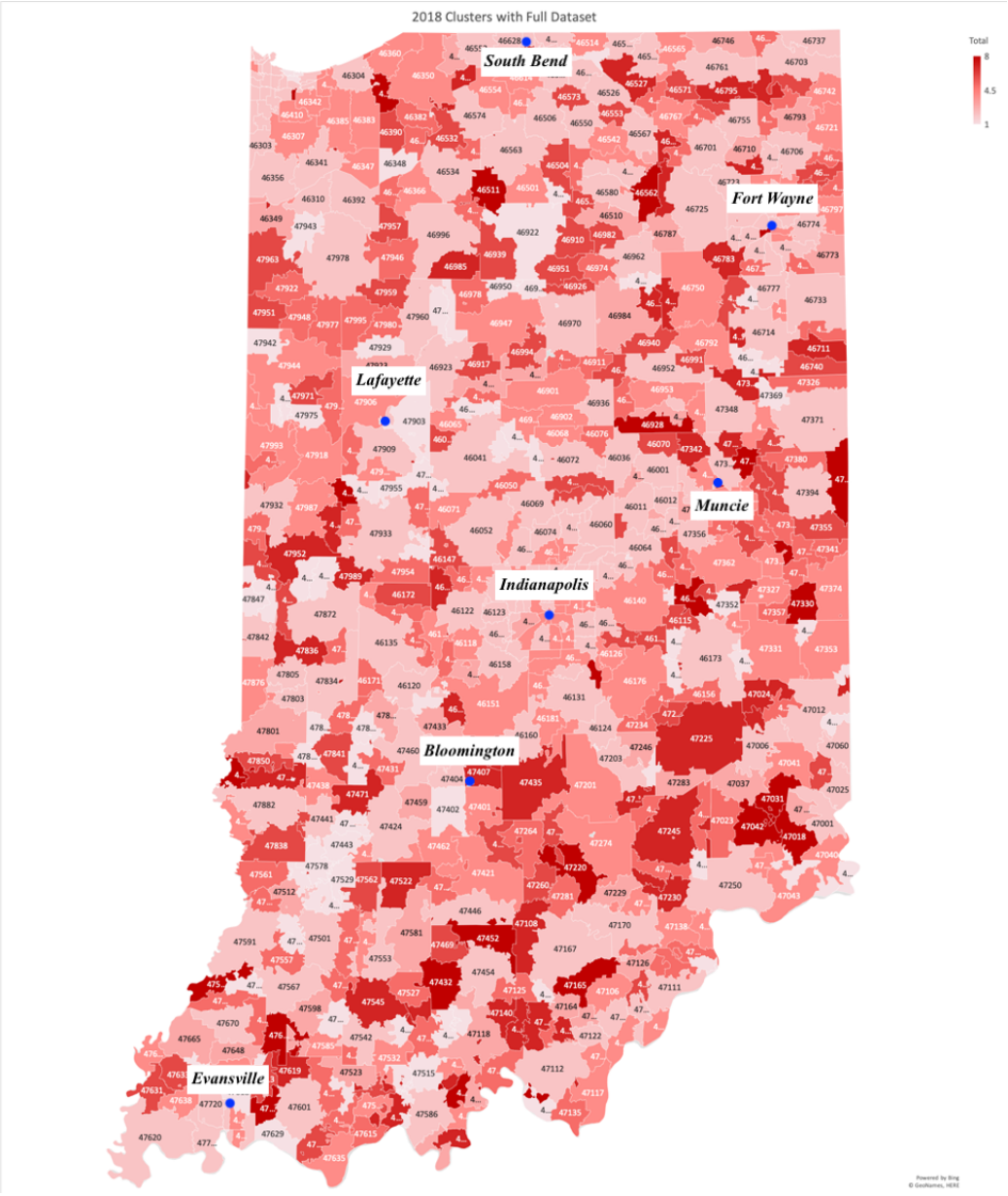

Figure 5.2: 2018 Cluster Map

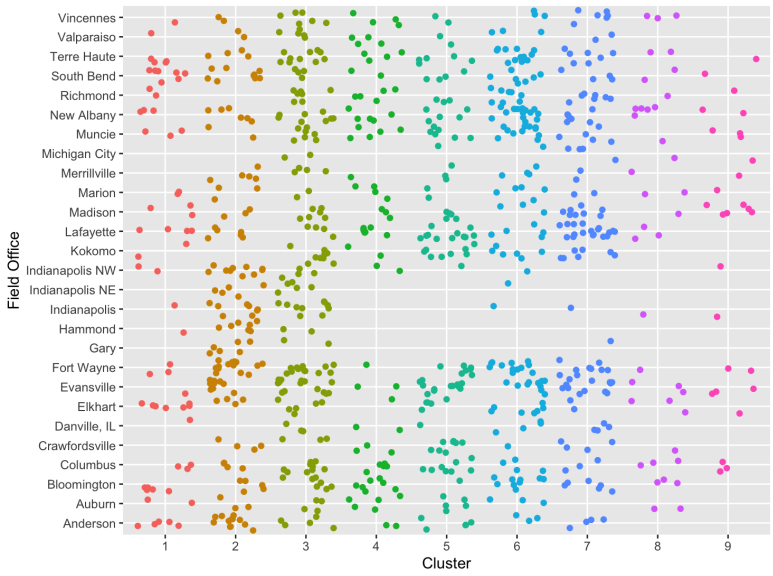

Figure 5.3: 2017 Field Offices by Cluster

these areas also belong to similar clusters in both years. In the maps, the ZIP codes located in the more rural parts of the state also belong to similar clusters.

Using this basic demographic information for the state of Indiana, the clustering appears to accurately capture the diversity of the state. Similar maps to the ones above for the models run with outliers removed can be found in the Appendix.

Figures 5.3 and 5.4 help to show the relationship between specific clusters and parts of the state of Indiana by showing what field offices are represented in each cluster. Given that Indianapolis is the state capital, it is logical that Indianapolis would differ from the rest of the state in the types of beneficiaries that live there. This is shown in Figure 5.3 for 2017, in which the majority of the ZIP codes for the Indianapolis field offices lie in the 2nd and 3rd clusters.

In Figure 5.4, a similar pattern is seen for the 2018 dataset, in which the Indianapolis ZIP codes lie primarily in the 2nd and 4th clusters. These figures also show the variability that exists in each cluster. As would be expected, ZIP codes that receive benefit payments from the same field office would not necessarily be similar and therefore would not be in the same cluster.

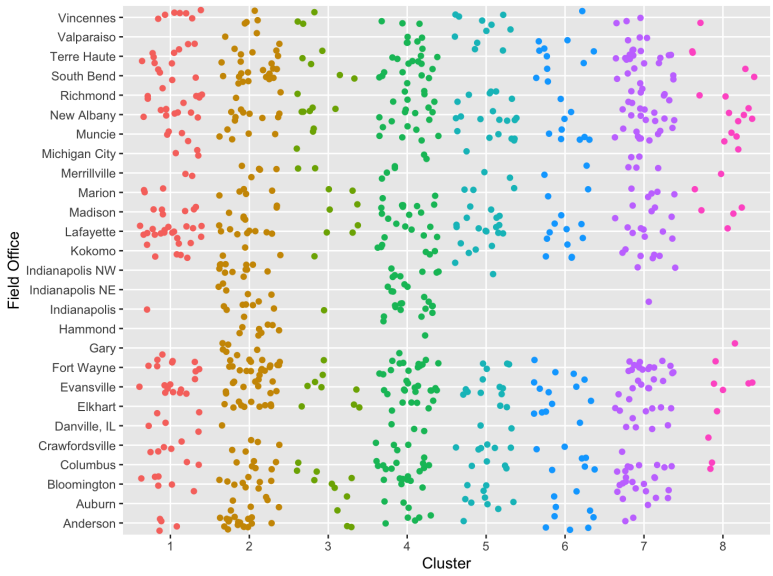

Figure 5.4: 2018 Field Offices by Cluster

5.1.1 2017 Clustering Results

Beyond the geographical information that can be explored from the clustering, there are many other unique results that resulted from the clustering of both datasets. One of the most unique results of the cluster analysis is the relationship between the number of specific types beneficiaries in a cluster and the amount of monthly benefit amounts paid out. It would be reasonably expected that clusters with a higher median number of beneficiaries would have a higher median value of monthly benefits paid. However, in a few cases, these are not the results seen in the anaylsis.

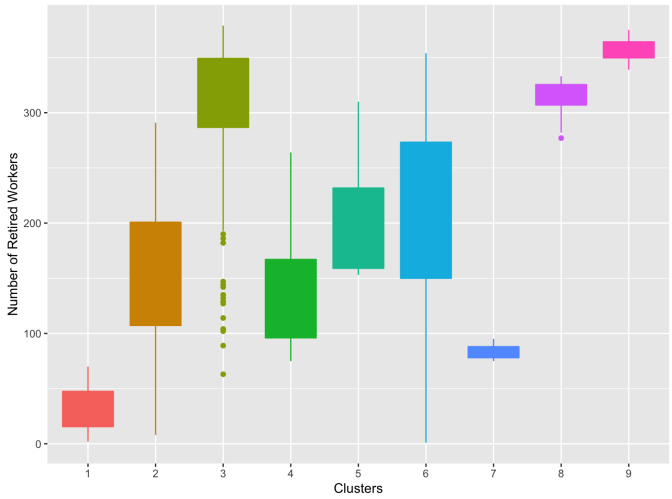

Figure 5.5: 2017 Retired Workers by Clusters

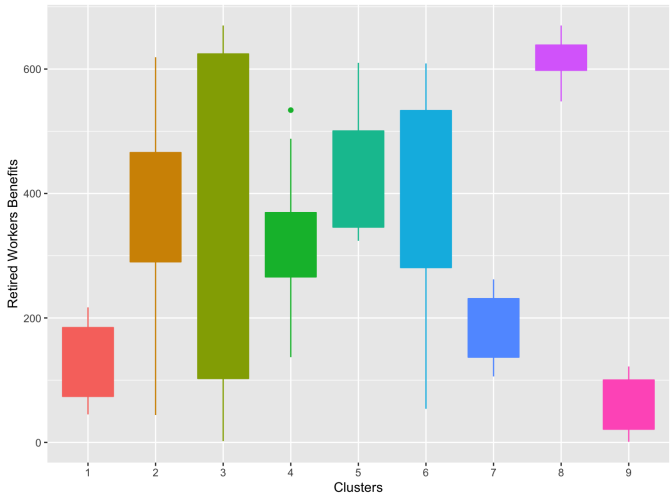

Figure 5.6: 2017 Retired Worker Benefits by Clusters

In Figure 5.5 and Figure 5.6 the boxplots for each clusters for the variables retired workers and retired worker monthly benefits are shown. While clusters 1, 2, 4, 5, 6, 7, and 8 have proportional medians for the two variables, clusters 3 and 9 do not. In cluster 3, the spread of retired workers is significantly larger than for the retired worker benefits. However, there is a large number of outliers for that particular cluster for retired workers. For cluster 9, the median number for retired workers is signficantly higher than for the median value for monthly benefits paid to retired workers.

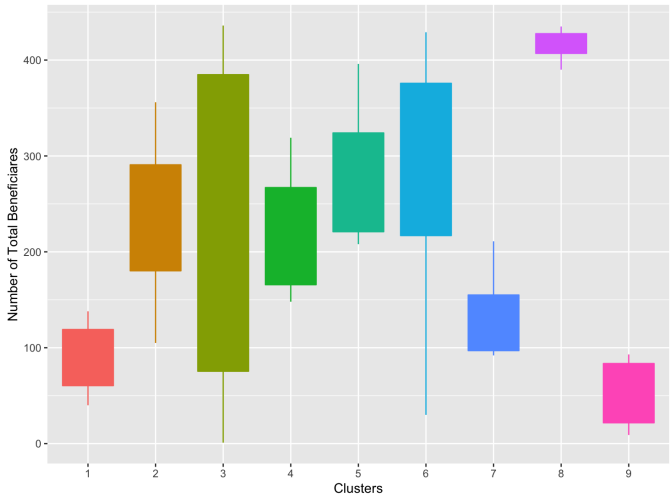

Figure 5.7: 2017 Total Beneficiaries by Clusters

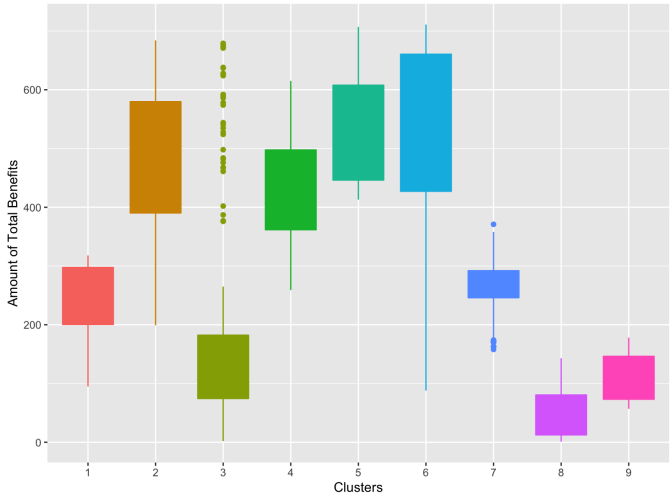

Figure 5.8: 2017 Total Monthly Benefits by Clusters

For the variables total beneficiaries and total monthly benefits, a similar relationship is found as with retired workers and retired workers benefits. In Figure 5.7 and Figure 5.8, the boxplots for the variables are shown by cluster. In cluster 3, the spread of the variable total monthly benefits is signifcantly smaller than for the variable total beneficiaries. Also, in cluster 8, the median value for total beneficiaries is larger than the other clusters. However, the median value for total monthly benefits for cluster 8 is smaller than the other clusters.

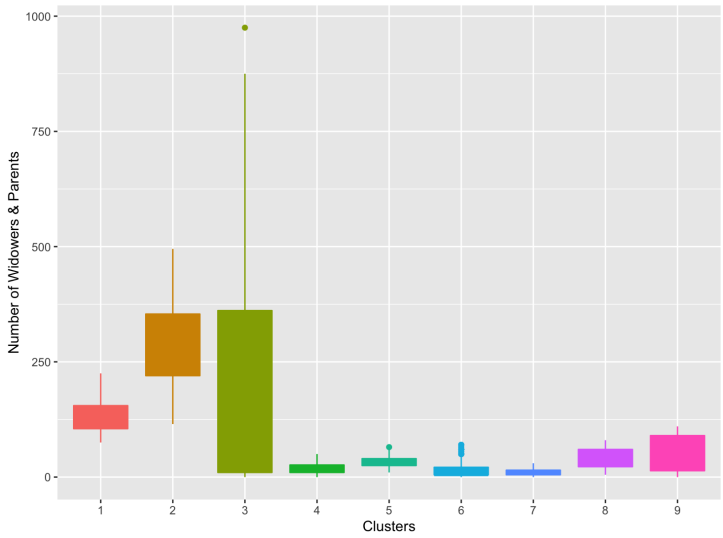

Figure 5.9: 2017 Widowers and Parents by Clusters

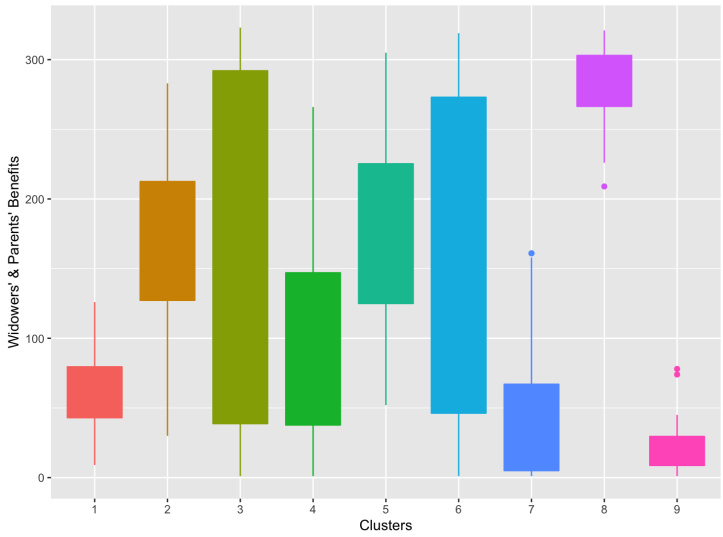

Figure 5.10: 2017 Widower and Parent Benefits by Clusters

As with the variables total beneficiaries and total monthly benefits, the variables widowers and parents and widower and parent benefits have unique relationships for the 3rd and 8th clusters. Shown in Figures 5.9 and 5.10 are the boxplots for the widowers and parents variables. For cluster 3, the spread of the variable widower and parent benefits is much larger than for the number of widowers and parents. For cluster 8, the median value for

widower and parent benefits is comparatively higher than the other clusters, when that is not the case for the median value for the number of widowers and parents.

Though there is a difference in spread in cluster 3 again, this and the other differences in spread can be explained by the inclusion of outliers in this dataset. As discussed previously in Figure 5.3, there was a large proportion of the Indianapolis ZIP codes in cluster 3. Because many of the Indianapolis ZIP codes were found to be outliers in the outlier analysis, it would make sense that they would be shown as outliers in these boxplots of the clusters. However, by including Indianapolis and the other outliers in this analysis, the whole state is able to be compared in the results.

5.1.2 2018 Clustering Results

Similar results to the 2017 cluster analysis can be found in the 2018 cluster analysis. Shown in Figure 5.11 and Figure 5.12 are the boxplots for the variables Retired Workers and Retired Worker Benefits by cluster. As shown in the graphs, clusters 1, 2, 3, 5, 6, and 7 seem to have median retired worker benefits that are similar to the median number of retired workers for the cluster.

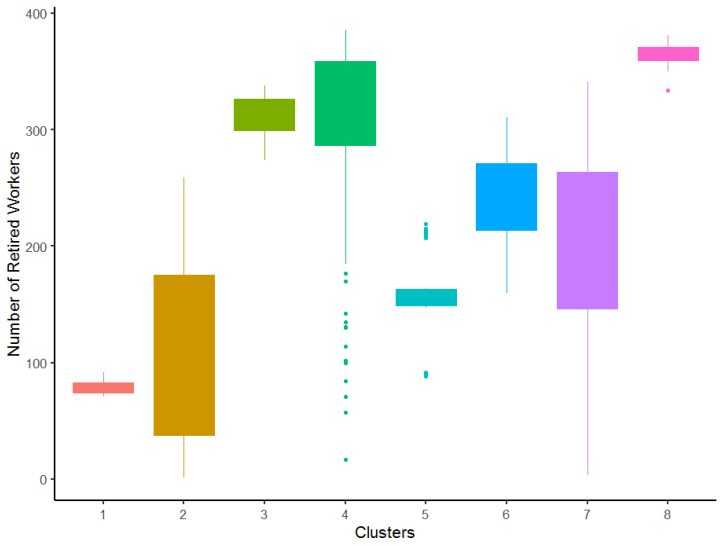

Figure 5.11: 2018 Retired Workers by Clusters

However, in cluster 4, the spread of the variable retired worker benefits is signifcantly larger than the spread of the number of retired workers. This

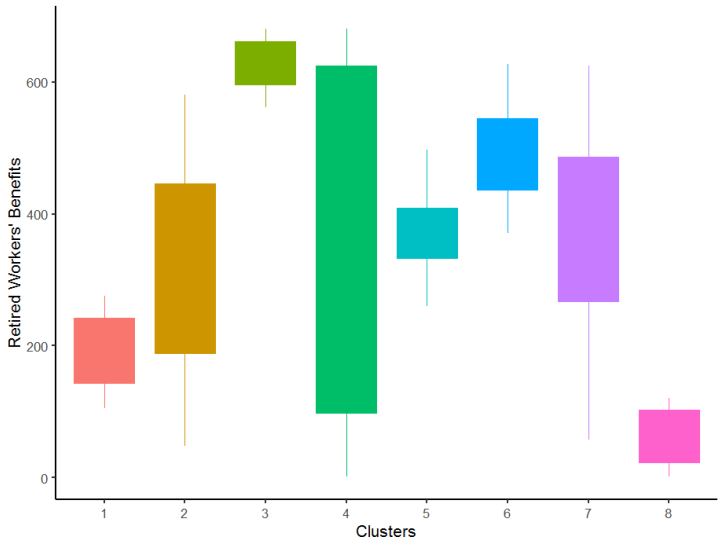

Figure 5.12: 2018 Retired Worker Benefits by Clusters

can be explained by the large number of outliers in cluster 4 for the variable number of retired workers. In cluster 8, the median number of retired workers is higher than the other clusters, but the median monthly benefits to retired workers is lower than the other clusters.

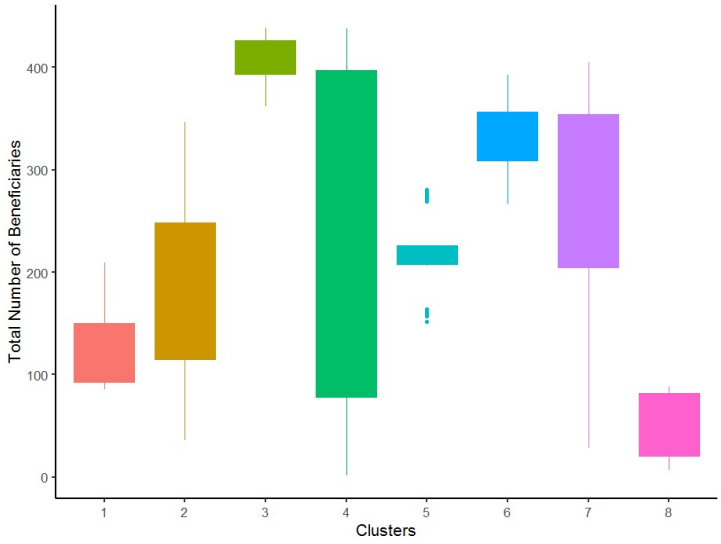

Figure 5.13: 2018 Total Beneficiaries by Clusters

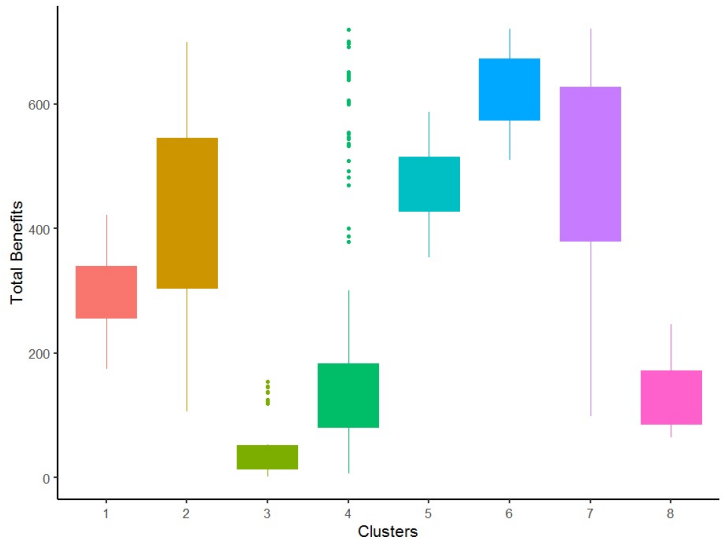

Figure 5.14: 2018 Total Monthly Benefits by Clusters

A similarly interesting relationship occurs with the variables total beneficiaries and total monthly benefits, as shown in Figure 5.13 and Figure 5.14. In the boxplots in the figures, the median values for clusters 1, 2, 5, 6, 7 and 8 are consistent between the two graphs, as would be expected. However, for cluster 3, the median value for the number of beneficiaires is significantly higher than the median amount of monthly benefits, compared to the other clusters. In cluster 4, the spread of the boxplot is larger for the number of total beneficiaries than for the amount of monthly benefits. However, this could be explained by the large number of outliers shown in the monthly benefits boxplot.

The variables widowers and parents and widower and parent benefits also showcased a unique relationship between the clusters. As shown in Figure 5.15 and Figure 5.16, the boxplots for clusters 3 and 7 vary from the pattern of the other boxplots. For cluster 3, the median number of widowers and parents eligible for benefits is comparatively lower than the median amount of benefits for the same group. In cluster 7, the spread of the amount of benefits for widowers and parents is much smaller than the spread for the number of widowers and parents.

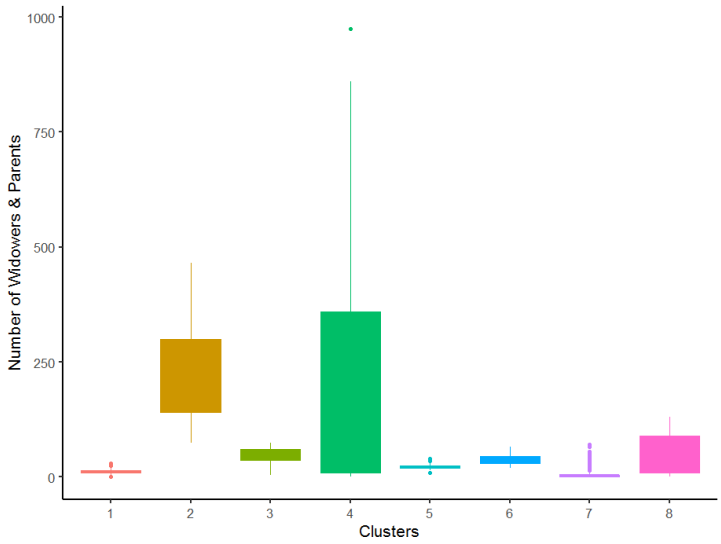

Figure 5.15: 2018 Widowers and Parents by Clusters

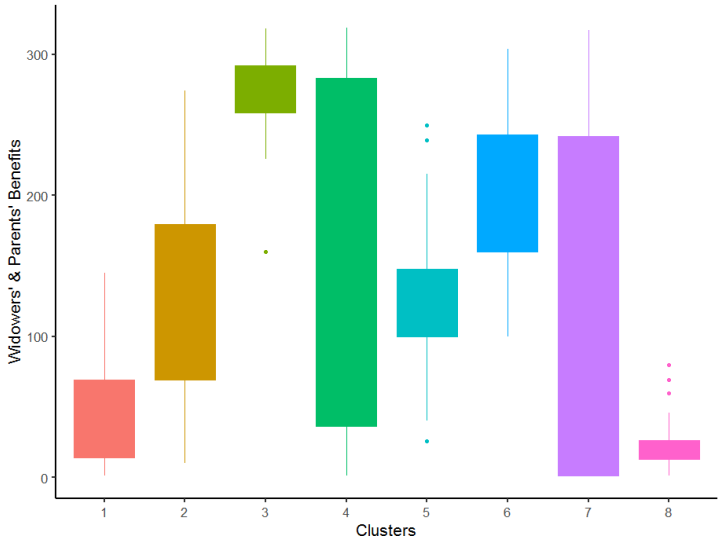

Figure 5.16: 2018 Widower and Parent Benefits by Clusters

5.2 Comparison of 2017 and 2018 Results

As discussed above, the results of the 2017 and 2018 models were unique for some specific clusters. In many cases, the median amount of benefits a particular group of beneficiaries was receiving did not align with the expectation

for the median number of beneficiaries in the area. Though it may be reasonably expected that the amount of monthly benefits an area receives would be consistent with the number of beneficiaries in the area, there are many complicating factors that go into determining how much money a beneficiary is entitled to receive. To see the all the median values for each variable for each cluster, refer to the Appendix.

According to the Social Security Administration retirement planner (2020), factors that affect the amount of benefits a retiree is eligible to receive include: receiving benefits while working; type of employment; government pension offset; income tax; maximum taxable earnings; income from pension, annuities, interest, and dividiends; Social Security Credits; and the Windfall Elimination Provision. The types of employment that impact the amount an individual is eligible to receive include: farm work, Federal Government Employment, household employment, military service, nonprofit or religious organization, railroad earnings, self-employment, state and local government employment, wages, work for a foreign government inside the U.S., and work outside the U.S.

With the datasets provided by the Social Security Administration for this research, the only information available is aggregated for the ZIP code in which benefits are received, so there is no information available about specific beneficiaries in the state of Indiana. However, it is reasonable to assume that the factors previously mentioned explain the clusters with a higher median number of retired workers and a lower median amount of monthly benefits.

Another aspect of the OASDI program that could explain the interesting relationships found is the survivors insurance portion of the program. The survivors insurance pays benefits to spouses, children, and parents in the event that a wage earner who paid into Social Security dies. These benefits are captured in the datasets used in this research in the variables, widow(er)s and parents, spouses and children.

According to the Social Security Administration survivors benefit planner (2020), the amount a survivor receives in benefits is based on the earnings of the wage earner who died. The monthly benefit amount is calculated as a percentage of the wage earner's basic Social Security benefit and many factors can affect the percentage a survivor is eligible to receive. Some of the factors include the age of the survivor, the disability status of the survivor, the sum of benefits received by all surviving family members, age a surviving spouse remaries, and the amount a survivor is currently earning. As with the retirement benefits, these factors could explain the clusters with variability in the median amount survivors received compared to the median number of survivors in the cluster.

Beyond the similar relationships that exist between specific variables and clusters in both models, there are other similarities between the two models. As would be expected, there is not a signficant change in the two datasets, as the individuals who are eligible for OASDI benefits are primarily the same from year to year, with the small differences due to the new individuals who become or elect to become eligible and to individuals who stop receiving benefits.

One large similarity, as discussed previously, it the treatment of ZIP codes in Indianapolis by both models. In the cluster analysis of both datasets, Indianapolis was mainly clustered into two clusters. Figures 5.1 and 5.2, which showcase the maps of the clustered ZIP codes, also display similarities between the two models. Many ZIP codes that are clustered together in 2017 remain clustered together in 2018. As can be seen by the maps of Indiana for the models when outliers removed, which are located in the Appendix, the outliers are also fairly consistent between the two datasets and are located in similar parts of the state.

Chapter 6

Discussion and Conclusion

6.1 Overview of Clustering Results

The OASDI program reports provide a unique opportunity for practical analysis of social insurance data. The goal of this research was to discover patterns in the OASDI benefits paid in the state of Indiana. This was accomplished through using model-based clustering to examine the data.

In order to use model-based clustering, maximum likelihood estimation was used to estimate parameter values for the Gaussian Mixture Model. The Expectation Maximization Algorithm was then employed to complete the maximum likelihood estimation, because there was missing information in the model. To determine the optimal number of clusters, Bayesian Information Criterion was utilized.

By selecting model-based clustering to analyze these datasets, the multimodal nature of the data was able to be utilized in finding patterns in the data. Beyond finding patterns in the data, the clustering was also successful in finding similarities between the benefits paid in 2017 and 2018, as well as minor differences.

After comparing the results of the 2017 and 2018 cluster analysis, a few conclusions could be developed. Since this research uses the two most recent reports from the OASDI program, the results of this analysis can be beneficial for considering trends in the state of Indiana over the past few years. The analysis also confirms that the OASDI program datasets are suitable for cluster analysis and that underlying patterns in benefits paid in the state of Indiana exist.

6.2 Conclusion

Overall, model-based clustering has shown to be an excellent data mining technique for finding patterns in Social Security data. This research found that the similar ZIP codes in Indiana tend to be ZIP codes that are similar in demographics and not necessarily in geographic locations. Also, over 2017 and 2018, the OASDI program did not change significantly in the state of Indiana. When considering other social insurance programs, model-based clustering is an effective technique for quickly identifying patterns based on quantitative data.

Though a variety of data mining techniques can be applied to the OASDI program reports, the ability of model-based cluster analysis used in this research to capture the multimodal nature of the reports is highly beneficial. This project aims to find the patterns that exist in the state of Indiana for the OASDI program over 2017 and 2018. Beyond finding numerous similarities between the two years, the clustering was able to identify regions of the state that receive lower monthly benefits than expected, due to a wide variety of regions. Similar techniques could prove helpful to governments and economists when examining the impact of social insurance programs.

Appendix A

Additional Figures and Tables

A.1 Cluster Maps for Models with Outliers Removed

Though this project ultimately decided to use the datasets that included outliers for analyzing the results of the clustering, the models with outliers removed still produced interesting maps of the clusters found. Shown in Figures A.1 and A.2 are the cluster maps for the datasets with outliers removed.

These graphs help to show the occurence of outliers around the state capital, Indianapolis, which is located in the center of the state. The consistency of the outliers between the two years can also be seen in the two maps.

A.2 Table of Median Values for Each Variable for Each Cluster

Shown in Tables A.1 and A.2 are the median variables for the majority of the variables in the datasets for each cluster. As explained in Chapter 5, the median monthly benefits of specific types of beneficiaries in some clusters is signficantly different than would be expected with the median number of beneficiaries for that particular cluster.

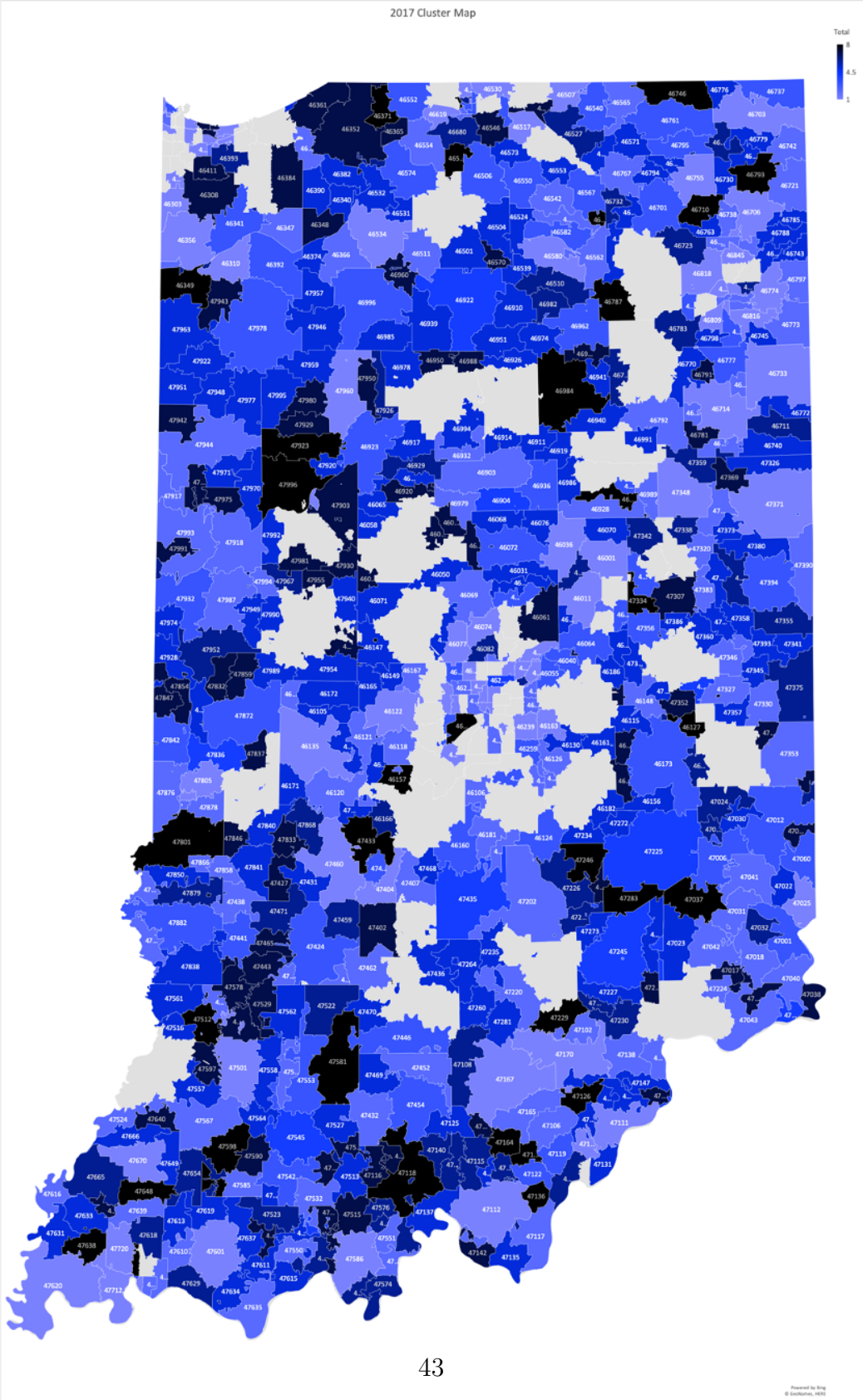

Figure A.1: 2017 Cluster Map without Outliers

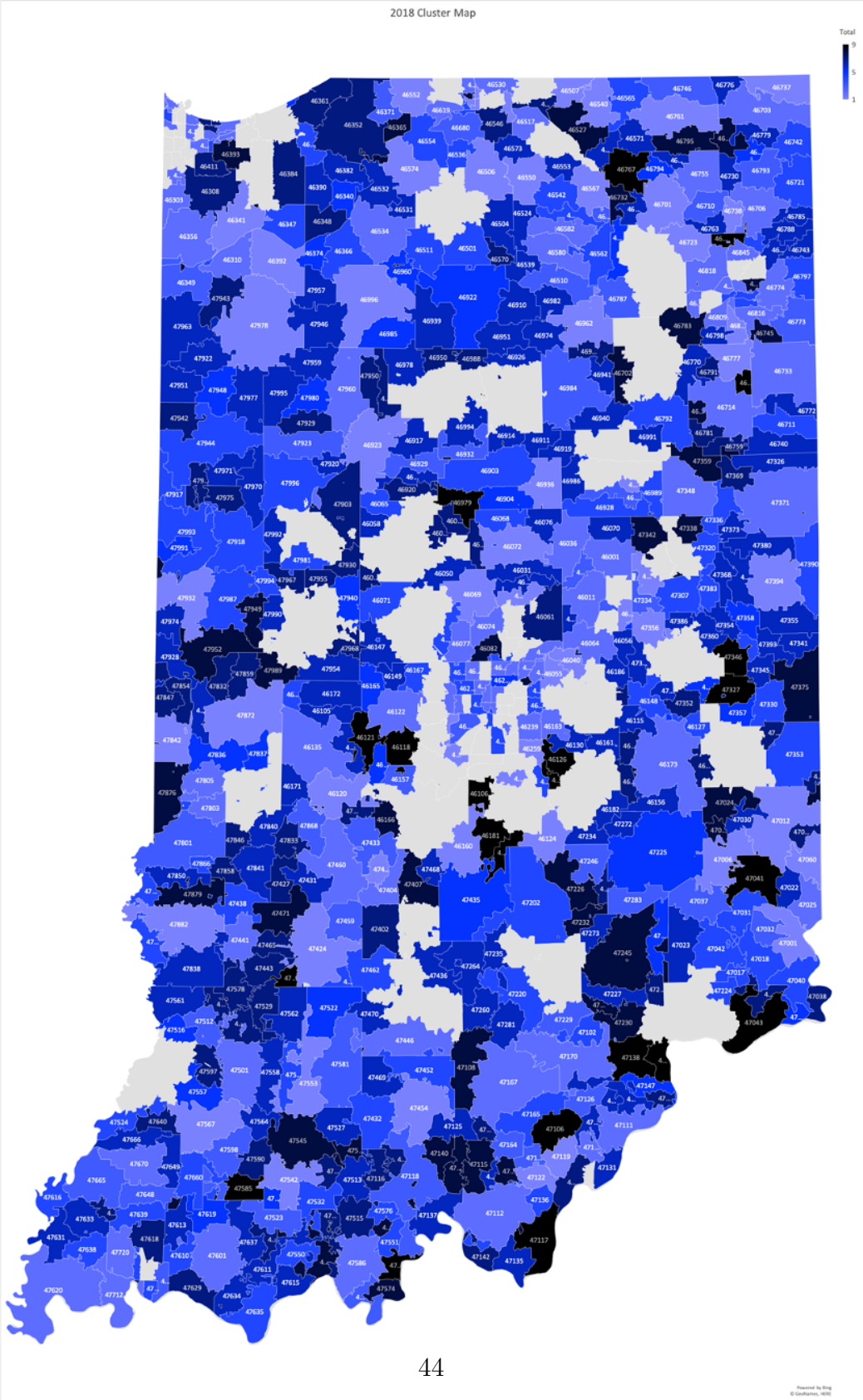

Figure A.2: 2018 Cluster Map without Outliers

	Disabled Workers	Retired Workers	Widowers & Parents	Spouses	Children	Retired Worker Benefits	Widowers & Spouses Benefits	Beneficiaries Over 65
1	78.0	33.0	125.0	65.0	130.0	167.0	58.0	39.0
2	140.0	142.0	280.0	132.5	295.0	380.5	168.5	166.5
3	59.0	330.5	70.0	40.0	90.0	452.5	210.5	323.0
4	122.0	153.0	20.0	10.0	20.0	282.5	109.0	149.5
5	147.0	217.5	30.0	20.0	30.0	433.5	174.0	214.0
6	46.0	222.0	5.0	5.0	5.0	427.5	206.0	212.0
7	95.0	84.0	10.0	5.0	10.0	154.0	43.0	81.0
8	51.0	309.0	50.0	25.0	55.0	627.0	287.0	314.0
9	63.0	355.5	77.5	40.0	72.5	35.5	17.5	370.0

Table A.1: 2017 Median Variable Values for Each Cluster

	Disabled Workers	Retired Workers	Widowers & Parents	Spouses	Children	Retired Worker Benefits	Widowers & Spouses Benefits	Beneficiaries Over 65
1	82.0	79.0	10.0	5.0	10.0	167.5	40.0	80.0
2	111.0	105.5	210.0	100.0	230.0	300.5	123.0	110.5
3	47.0	308.0	50.0	25.0	55.0	631.0	280.0	308.5
4	53.0	325.0	65.0	35.0	70.0	467.0	229.0	312.0
5	120.0	155.0	20.0	10.0	20.0	356.0	126.0	154.0
6	161.0	224.0	35.0	20.0	35.0	501.0	206.0	254.0
7	50.0	158.0	5.0	0.0	5.0	369.0	159.0	198.0
8	61.0	363.0	80.0	32.5	72.5	36.5	18.0	373.0

Table A.2: 2018 Median Variable Values for Each Cluster

Bibliography

Ahlquist, John S., and Christian Breunig. "Country Clustering in Comparative Political Economy." No. 09/5. MPIfG discussion paper, 2009.

"Benefits Planner: Retirement." *Social Security*, Social Security Administration, 2020, www.ssa.gov/planners/retire/.

"Benefits Planner: Survivors." *Social Security*, Social Security Administration, 2020, www.ssa.gov/planners/survivors/.

Diamond, Peter A. "A framework for social security analysis." *Journal of Public Economics* 8.3 (1977): 275-298.

Feldstein, Martin S, and Jeffrey B Liebman. *The Distributional Aspects of Social Security and Social Security Reform*. University of Chicago Press, 2002. Accessed 26 Feb. 2020.

Feldstein, Martin, and Jeffrey B. Liebman. "Social Security." *Handbook of public economics* 4 (2002): 2245-2324.

Fraley, Chris, and Adrian E. Raftery. "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis." The computer journal 41.8 (1998): 578-588.

Gough, Ian. "Social Assistance Regimes: A Cluster Analysis." *Journal of European social policy* 11.2 (2001): 165-170.

KaewTraKulPong, Pakorn, and Richard Bowden. "An Improved Adaptive Background Mixture Model for Real-Time Tracking with Shadow Detection." *Video-based surveillance systems*. Springer, Boston, MA, 2002. 135-144.

Klugman, Stuart A., et al. *Loss Models: From Data to Decisions*. John Wiley & Sons. 4th Edition. 2012.

Meseguer, Javier. "Outcome Variation in the Social Security Disability Insurance Program: The Role of Primary Diagnoses." *Soc. Sec. Bull.* 73 (2013): 39.

Miljkovic, Tatjana, and Bettina Grün. "Modeling Loss Data using Mixtures of Distributions." *Insurance: Mathematics and Economics* 70 (2016): 387-396.

"OASDI Beneficiaries by State and ZIP Code, 2017." Social Security, Social Security Administration, July 2019.

"OASDI Beneficiaries by State and ZIP Code, 2018." Social Security, Social Security Administration, July 2019.

Povey, Daniel, et al. "The Subspace Gaussian Mixture Model –A Structured Model for Speech Recognition." *Computer Speech & Language* 25.2 (2011): 404-439.

Scrucca, Luca. "Dimension Reduction for Model-Based Clustering." Statistics and Computing 20.4 (2009): 471–484. Crossref. Web.

Tan, Pang-Ning, et al. "Cluster Analysis: Basic Concepts and Algorithms." *Introduction to Data Mining.* Pearson. 2006. pp. 487–568.